7

Computational SNP Discovery in DNA Sequence Data

Gabor T. Marth

1. Introduction

Both the quantity and the distribution of variations in DNA sequence are the product of fundamental biological forces: random genetic drift, demography, population history, recombination, spatial heterogeneity of mutation rates, and various forms of selection. In humans, single base-pair substitution-type sequence variations occur with a frequency of approx 1 in 1.3 kb when two arbitrary sequences are compared ***(1)***. This frequency increases with higher sample size ***(2)***, i.e., we expect to see, on average, more single nucleotide polymorphisms (SNPs) when a higher number of individual chromosomes are examined ***(3,4)***.

SNPs currently in the public repository ***(5)*** were discovered in DNA sequence data of diverse sources, some already present in sequence databases, but the majority of the data generated specifically for the purpose of SNP discovery. Nearly 100,000 SNPs in transcribed regions were found by analyzing clusters of expressed sequence tags (ESTs) ***(6–8)***, or by aligning ESTs to the human reference sequence ***(9)***. The three major sources of genomic SNPs were sequences from restricted genome representation libraries ***(10)***, ran-

From: *Methods in Molecular Biology, vol. 212:*
Single Nucleotide Polymorphisms: Methods and Protocols
Edited by: P-Y. Kwok © Humana Press Inc., Totowa, NJ

dom shotgun reads aligned to genome sequence ***(1)***, and the overlapping sections of the large-insert (mainly bacterial artificial chromosome, or BAC) clones sequenced for the construction of the human reference genome ***(11–13)***. Most of these SNPs were detected in pairwise comparisons where one of the two samples was a genomic clone sequence. Theory predicts ***(14)***, and experiments confirm, that shallow sampling results in an overrepresentation of common variations: these common SNPs tend to be ancient variations, often present in all or most human populations ***(15)*** and expected to be valuable for detecting statistical association ***(16)***. For the same reason, many rare polymorphisms with rare phenotypic effects are likely to be absent from this set. The current collection of SNPs forms a dense, genome-wide polymorphism map ***(1)*** intended as a starting point for regional variation studies. An exhaustive survey of polymorphisms in a given region of interest is likely to require significantly higher sample sizes. Even so, the isolation of rare phenotypic mutations may only be possible by the crosscomparison between large samples of affected patients and those of controls.

Computational SNP discovery, in a general sense, refers to the process of compiling and organizing DNA sequences that represent orthologous regions in samples of multiple individuals, followed by the identification of polymorphic sequence locations. The first step typically involves a similarity search with the Basic Local Alignment Search Tool (BLAST) ***(17)*** to compile groups of sequences that originate from the region under examination. This is followed by the construction of a base-wise multiple alignment to determine the precise, base-to-base correspondence of residues present in each of the samples in a group. Finally, each position of the multiple alignment is scanned for nucleotide mismatches.

Some of the most serious difficulties of sequence organization stems from the repetitive nature of the DNA observed in many organisms. It is well known that nearly half of the human genome is made up of high copy-number repetitive elements ***(18,19)***. In addition, many intra- and interchromosomal duplication exist, a large number of them yet uncharacterized. Similar to members of multigene families, these duplicated (paralogous) genomic regions may exhibit extremely high levels of sequence similarity ***(18)***, sometimes over 99.5%, and can extend over hundreds of kilobases. Failure to distinguish between sequences from different copies of duplicated regions results in false SNP predictions that represent paralogous sequence differences rather than true polymorphisms.

The construction of correct base-wise multiple alignments is a difficult problem because of its computational complexity. Sequences under consideration are generally of different length rendering global sequence alignment algorithms such as CLUSTALW ***(20)*** rarely applicable. Expressed sequences (ESTs or more or less complete gene sequences) require local alignment techniques that are unperturbed by exon-intron punctuation and alternatively spliced sequence variants.

Once a multiple alignment is constructed, nucleotide differences among individual sequences can be analyzed. Owing to the presence of sequencing errors, not every nucleotide position with mismatches automatically implies a polymorphic site. Although it is impossible to decide which is the case with certainty, the success of SNP detection ultimately depends on how well one is able to discriminate true polymorphisms from likely sequencing errors. This is usually accomplished by statistical considerations that take advantage of measures of sequence accuracy ***(21,22)*** accompanying the analyzed sequences. The result, ideally, is a set of candidate SNPs, each with an associated SNP score that indicates the confidence of the prediction. Accurate confidence values can be extremely useful for the experimentalist in selecting which SNPs to use in a study or for further characterization, and enables one to use the highest number of candidates within the bounds of an acceptable false positive rate.

2. Materials

Sequences used in SNP analysis come from diverse sources. From the viewpoint of sequence accuracy, they can be categorized as either single-pass sequence reads or consensus sequences that result from multipass, redundant sequencing of the same underlying DNA.

The overall sequencing error rate of single-pass sequences is in the 1%-range ***(21–23)***, an order of magnitude higher than the average polymorphism rate (roughly 0.1%). The error rate is typically much higher at the beginning and the end of a read ***(21,22)***. Clusters of sequencing errors are also common; the location of these is highly dependent on specific base combinations, as well as the sequencing chemistry used. For detecting sequence variations, even marginally accurate data can be useful as long as regions of low accuracy nucleotides can be avoided. The most widely used base-calling program, PHRED ***(21,22)*** associates a base quality value to each called nucleotide. This base quality value, Q, is related to the likelihood that the nucleotide in question was determined erroneously: $Q = -10 \log_{10}(P_{error})$. Although different sequencing chemistries pose different challenges to base calling, tests involving large data sets have demonstrated that the quality value produced by PHRED is a very good approximation of actual base-calling error rates ***(21,22)***. Using base quality values, mismatches between low-quality nucleotides can be discarded as likely sequencing errors. Because consensus sequences are the product of multiple sequence reads, they are generally of higher accuracy. Exceptions to this rule are regions where the underlying read coverage is low, and/or regions where all underlying reads are of very low quality. Recognizing this problem, sequence assemblers (computer programs that create consensus sequences) also provide base quality values for the consensus sequence by combining quality scores of the underlying reads ***(24,25)***. The following subsections describe the most commonly used sequence sources used in SNP discovery.

2.1. STS Sequences

Sequence-tagged site (STS) sequences, amplified and sequenced in multiple individuals, were used in the first large-scale efforts to catalog variations at the genome scale ***(26)***. One of the main advantages of this strategy was that PCR primers, optimized during STS development, were readily available for use. If starting material for the amplification is genomic DNA, these sequences represent the superposition of both copies of a chromosome within an individual. As a result, the sequence may contain nucleotide ambiguities that correspond to heterozygous positions in the individual. Base-calling algorithms trained for homozygous reads will assign a low base quality value to whichever nucleotide is called, rendering base quality value-based SNP detection algorithms ineffective for these reads. Specialized algorithms ***(31)*** have been designed to deal with heterozygote detection, as discussed next.

2.2. EST Sequences

Expressed Sequence Tag (EST) Reads represent the richest source of SNPs in transcribed regions ***(6–8,27,28)*** to date. The majority of ESTs are single-pass reads, often from tissue-specific cDNA libraries ***(29,30)***. Because a single EST read may contain several exons, special care must be taken when these reads are aligned to genomic sequences. An additional difficulty is the alignment of ESTs representing alternative splice-variants of a single gene.

2.3. Small Insert Clone Sequences

2.3.1. Sequences from Reduced Representation Libraries

Size-Selected Restriction Fragments recognized by specific restriction enzymes are quasirandomly distributed in genomic DNA. The average distance between neighboring restriction sites (restriction fragment length) is a function of the length of the recognition sequence. A reduced, quasirandom representation of the genome can be achieved by first constructing a library of cloned restriction fragments, followed by size-selection to exclude fragments outside a desired length range. The number of different fragments (complexity) present in the library can be precalculated for any given length range. Inversely, library complexity can be controlled by appropriate selection of the upper and lower size limits ***(10)***.

2.3.2. Sequences from Random Genomic Shotgun Libraries

Random Genomic Subclone Reads are sequenced from DNA libraries with a quasirandom, short-insert subclone representation of the entire genome (whole-genome shotgun libraries). Because these reads deliver a random sampling of the whole genome, they are well-suited for genome-wide SNP discovery ***(1,12)***.

2.4. Large-Insert Genomic Clone Consensus Sequences

Recent large-scale, genome-wide SNP discovery projects ***(1,11–13,32)*** take advantage of the public human reference sequence built as a tiling path through partially overlapping, large-insert genomic clones ***(18,23)***. The sequence of these clones was determined with a local shotgun strategy. By cloning random fragments into a suitable sequencing vector, a subclone library is created for each clone. This library is then extensively sequenced until reaching a desired, three- to tenfold, quasirandom read coverage. The DNA sequence of the large-insert clone is reconstructed by assembling the shotgun reads with computer programs ***(24)***. At this stage, there are still several gaps in the sequence, although overall accuracy is high (approx 99.9%). Gap closure and clean up of regions of low-quality sequence requires considerable manual effort ***(23)*** known as “finishing.” Finished or “base-perfect” sequence is assumed at least 99.99% accurate ***(18)***.

2.5. Assembled Whole-Genome Shotgun Read Consensus Sequences

Similar in nature to genomic clone sequences, these consensus sequences are the result of assembling a large number of genome-wide shotgun reads, possibly from libraries representing multiple individuals. Over two million human SNP candidates were discovered in the private sector by the analysis of multi-individual reads that provided the raw material for the construction of a human genome reference sequence produced by the whole-genome sequence assembly method ***(19)***.

3. Methods

3.1. Published Methods of SNP Discovery

Methods of SNP mining have gone through a rapid evolution during the past few years. The first approaches relied on visual comparison of sequence traces from multiple individuals ***(33)***. Although manual comparison of a small number of sequence traces is feasible, standard accuracy criteria are hard to establish, and this method does not scale well for multiple sequence traces and many polymorphic locations. The efficiency of visual inspection is increased when it is performed in the context of a multiple sequence alignment ***(27,34,35)***, aided by computer programs that are capable of displaying the alignments and provide tools for simultaneous viewing of sequence traces at a given locus of the multiple alignment ***(36)***. Computer-aided prefiltering followed by manual examination of sequence traces ***(11,32)*** was used in the analysis of overlapping regions of genomic clone sequences to detect candidate SNPs as sequence differences between reads representing the two overlapping clones. These early methods were instrumental in demonstrating the value of extant sequences, sequenced as part of the Human Genome Project, for the discovery of DNA sequence variations. Although visual inspection remains an integral part of software testing and tuning, demands for fast and reliable SNP detection in large data sets have necessitated the development of automated, computational methods of SNP discovery.

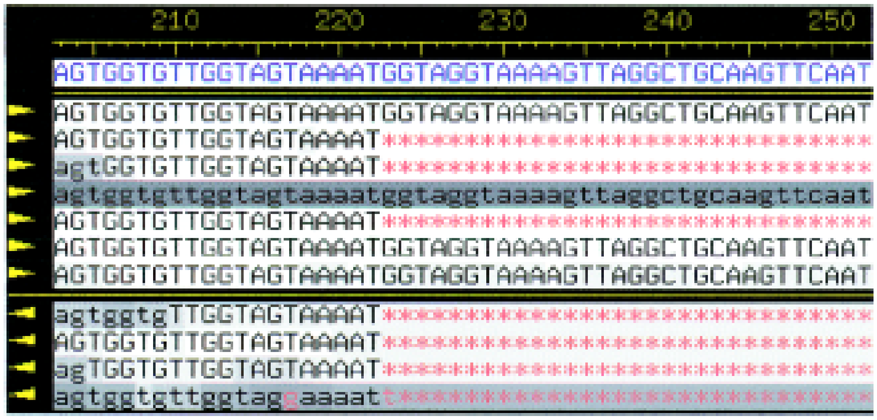
The first generation of these methods was designed to enable mining the public EST database ***(37)***, and relied, in part, on tools previously developed to aid the automation of DNA sequencing ***(23)***. SNP detection was performed by software implementing heuristic considerations. Picoult-Newberg et al. ***(27)*** used the genome fragment assembler PHRAP to cluster and multiply align ESTs from 19 cDNA libraries. The use of the genome assembler implied that alternatively spliced ESTs were not necessarily included in a single cluster. There was no attempt to distinguish between closely related members of gene families (paralogs). SNP detection was carried out through the successive application of several filters to discard SNP candidates in low-quality regions, followed by manual review. Mainly as the result of conservative heuristics, this method only found a small fraction, 850 SNP candidates in several hundreds of thousands of sequences analyzed. Buetow et al. ***(6)*** used UNIGENE ***(38)***, a collection of precomputed EST clusters as a starting point. ESTs within each cluster were multiply aligned with PHRAP ***(24)***. Identification of paralogous subgroups within clusters was done by constructing phylogenetic trees of all cluster members and analyzing the resulting tree topology. Again, SNP candidates were identified by heuristic methods to distinguish between true sequence differences and sequencing errors. This method yielded over 3,000 high-confidence candidates in 8,000 UNIGENE clusters that contained at least 10 sequence members. Unfortunately, the great majority of clusters contained significantly fewer sequences that could not be effectively analyzed with these methods.

The development of a second generation of tools was prompted by the needs of genome-scale projects of SNP discovery. The large amount of data generated by The SNP Consortium (TSC) ***(1)*** has spurred the development of several SNP discovery tools. In the initial phase, the TSC employed a molecular strategy called restricted genome representation (RRS), which involves the sequencing of size-selected restriction fragment libraries from multiple individuals ***(10)***. For example, the full digestion by a given restriction enzyme may produce 20,000 genomic fragments in the 450–550-bp length range. After digestion of the genomic DNA of each of the 24 individuals, followed by size-selection, the restriction fragment libraries are pooled. When a collection of such random fragments is sequenced to appreciable redundancy (say, 60,000–80,000 reads), the sequence of many of the fragments will be available from more than one individual. These redundant sequences are a suitable substrate for SNP analysis. The analysis of data of this type is similar to that of EST sequences. First, one must cluster the sequence reads to delineate groups of identical fragments. To avoid grouping sequences based on similarity between known human repeats they contain, the reads are screened and repetitive sequences are masked ***(39)***. Pairs of similar sequences are determined by a full pair-wise similarity search between all reads from a given library. Pairs are merged into groups (cliques) by single-linkage, transitive clustering. Some groups may still be composed of sequences that represent low-copy repeats (paralogous regions) not present in the REPEATMASKER repeat-sequence library. One of the strategies to identify these potential paralogs is to compare cluster depth (the number of sequences in the group) to expectations obtained from Poisson sampling with the given redundancy ***(10)***. Groups that survive these filtering steps are analyzed for SNPs. One of the methods used is based on establishing a quality standard for each of the aligned nucleotides within each sequence, taking into account the base quality value of the nucleotide in question as well as the quality of the neighboring nucleotides ***(10***; Neighborhood Quality Standard, or NQS). Instead of the full multiple alignment, the detection of SNPs was based on the analysis of all possible read pairs within a given group: mismatches between pairs of aligned nucleotides meeting the NQS were extracted as SNP candidates.

As the initial, draft sequencing of the human genome neared completion, it was possible to switch towards a more accurate, more efficient strategy. As the majority of the genome was available as genome reference sequence ***(18)***, sequencing of whole-genome, random, subclone libraries would provide sequence coverage that could be compared to the reference sequence. This reduced the time and cost associated with the creation of restricted representation subclone libraries ***(10,18)***. The informatics problems associated with this strategy were also reduced in complexity. It was now possible to use a single similarity search to place the fragments on the genome reference. By the same procedure, it was also possible to ascertain alternative (paralogous) locations. This is the strategy employed by the algorithm SSAHASNP ***(40)***, which combines a fast search algorithm of short-sequence fragments against the genome with a SNP detection algorithm that uses the NQS ***(10)*** to find SNP candidates in pair-wise comparisons of sequence fragments against the genome. As a fast tool capable of efficient processing of large data sets, SSAHASNP was used in the discovery of a large fraction of SNPs in the TSC data ***(1)***.

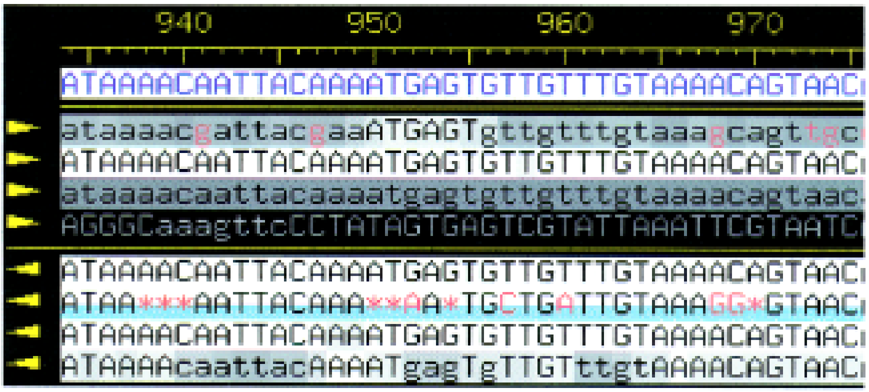
As we can see from the previous discussion, the molecular substrates involved in different projects of sequence-based SNP discovery represent data of varied types and sequence sources. The result is a multitude of different scenarios in terms of alignment depth, what the individual sequences represent, overall sequence accuracy, and so on. The methods of SNP discovery we have discussed so far are generally quite successful in operating within the specific sequence context for which they were developed. There was, however, a growing need for general tools of SNP discovery ***(41)*** that are able to analyze sequences both in shallow or in deep coverage, sequences of different sources simultaneously, without human review, and assign a realistic measure of confidence in the SNP candidates, without regard to the source and overall accuracy of these sequences. To achieve the flexibility this required, it was necessary to develop mathematically rigorous, statistical methods of SNP detection. Here we will describe POLYBAYES ***(9)***, one of the first general-purpose SNP analysis tools available for use today.

POLYBAYES is composed of three parts, each independent of the others: an anchored multiple alignment algorithm, a paralog discrimination algorithm, and the SNP detection algorithm. The anchored alignment algorithm assumes the availability of a genomic reference sequence (such as the Genome Assembly ***[18]*** for the Human Genome). Short-sequence fragments are organized by aligning them to the reference sequence. This algorithm works well in the case of cDNA (EST) sequences even in the presence of alternative splicing, as individual exons are aligned while leaving gaps for the introns or spliced-out exons (*see* **Fig. 1**). The paralog discrimination algorithm examines the alignment of the fragment to the genomic reference, and decides, on the basis of the sequence quality information, whether the number of discrepancies observed in the alignment is statistically consistent with the number expected from polymorphisms plus sequencing errors. If the number of observed discrepancies greatly exceeds the number expected, the



```
     210       220       230       240       250
AGTGGTGTTGGTAGTAAAATGGTAGGTAAAAGTTAGGCTGCAAGTTCAAT
AGTGGTGTTGGTAGTAAAATGGTAGGTAAAAGTTAGGCTGCAAGTTCAAT
AGTGGTGTTGGTAGTAAAAT******************************
agtGGTGTTGGTAGTAAAAT******************************
agtggtgttggtagtaaaatggtaggtaaaagttaggctgcaagttcaat
AGTGGTGTTGGTAGTAAAAT******************************
AGTGGTGTTGGTAGTAAAATGGTAGGTAAAAGTTAGGCTGCAAGTTCAAT
AGTGGTGTTGGTAGTAAAATGGTAGGTAAAAGTTAGGCTGCAAGTTCAAT
agtggtgTTGGTAGTAAAAT******************************
AGTGGTGTTGGTAGTAAAAT******************************
agTGGTGTTGGTAGTAAAAT******************************
agtggtgttggtaggaaaatt*****************************
```

Fig. 1. Alignment of EST reads to genomic anchor sequence (viewed in the CONSED sequence viewer-editor program). ESTs in this alignment represent two alternative splice variants, both correctly aligned to the genome sequence.



```
  940       950       960       970
ATAAAACAATTACAAAATGAGTGTTGTTTGTAAAACAGTAAC
ataaaacgattacgaaATGAGTgttgtttgtaaagcagttgc
ATAAAACAATTACAAAATGAGTGTTGTTTGTAAAACAGTAAC
ataaaacaattacaaaatgagtgttgtttgtaaaacagtaac
AGGGCaaagttcCCTATAGTGAGTCGTATTAAATTCGTAATC
ATAAAACAATTACAAAATGAGTGTTGTTTGTAAAACAGTAAC
ATAA***AATTACAAA**AA*TGCTGATTGTAAAGG*GTAAC
ATAAAACAATTACAAAATGAGTGTTGTTTGTAAAACAGTAAC
ATAAAAcaattacAAAATgagTgTTGTttgtAAAACAGTAAC
```

Fig. 2. Example of a paralogous EST sequence (marked with blue bar) in alignment with sequences likely to originate from the given genomic locus. The paralog is detected and tagged automatically by the software.

sequence fragment is flagged as a likely paralog, and is discarded from further analysis (*see* **Fig. 2**).

The SNP detection algorithm employed by POLYBAYES calculates the probability that discrepancies at the analyzed location represent true sequence variation as opposed to sequencing error. As a

Bayesian algorithm, it combines *a priori* (prior) knowledge about the sequence context with the specific, observed data represented by the sequences under examination. Typically, such prior knowledge includes an approximate average polymorphism rate in the region, and the expected ratio between transitions and transversions. Additional information may include the knowledge of the number of different individuals represented by the sequences within the alignment, or the degree of their relatedness. Often, multiple sequence reads (e.g., forward-reverse read pairs) may originate from a single DNA clone template; in such cases, any mismatch between these reads is *a priori* identified as a sequencing error. The role of sequence accuracy, as expressed by the base quality values in the individual sequences, is quite intuitive: a mismatch between nucleotides of low accuracy is more likely the result of sequencing error than that of true variation. On the other hand, if a mismatch occurs between nucleotides with high base quality values, the likelihood of a true polymorphism is higher. Alignment depth (the number of sequences contributing to the site under examination) is similarly important: a candidate A/G polymorphism between only two sequences may be less convincing than in a situation where, say 30 sequences contribute an A and another 30 sequences contribute a G residue to the alignment slice. Finally, the effect of base compositional biases may be significant in extremely A/T or G/C rich organisms, and is taken into account in the computations. The algorithm can be summarized as follows: At a given slice of N aligned nucleotide sequences, each sequence can represent one of the four DNA nucleotides, giving rise to a total of 4^N possible permutations within the slice. The POLYBAYES algorithm calculates the Bayesian posterior probability for all 4^N possible permutations taking into account the prior expectations, the base quality values, local base composition, and the alignment depth. The sum of the probabilities for all polymorphic permutations (i.e., permutations whereby not all N sequences are in agreement) is the likelihood that the sequences at the given location harbor a SNP. Because the algorithm does not depend on the source of the quality values (whether generated by a base caller such as PHRED, or by a fragment assembly program such as PHRAP) it is possible to objectively and simultaneously evaluate all available data present in the alignment, without regard to sequence source or restrictions on data quality. For each site of the alignment, the algorithm outputs the probability that the site is polymorphic. These probability values were shown to accurately estimate the validation rate of candidate SNPs in various mining applications ***(1,9,15)***. This is desirable because realistic estimates for the true positive rate allow one to use the highest number of SNP candidates within an acceptable false positive rate. The POLYBAYES software is compatible with the PHRED/PHRAP/CONSED file structure, is capable of analyzing multiple alignments created with PHRAP, and the output, including markup information such as paralog tags and candidate SNP sites, is directly viewable within CONSED (**Figs. 2** and **3**). An alternative statistical formulation ***(8)*** developed to analyze EST clusters produces a log-odds (LOD) score to rank SNP candidates based on sequence accuracy, the quality of the alignment, prior polymorphism rate, and by evaluating adherence to the rules of Mendelian segregation of alleles within individual cDNA libraries.

There are two additional cases of practical importance that the algorithms described earlier were not designed to work with directly. In many situations, the DNA template that is available for analysis is double stranded, genomic DNA of an individual, or sometimes a pool of multiple individuals. The first is the case when a known region is assayed from the genomic DNA of multiple individuals ***(34,35)***, giving rise to sequence traces that contain heterozygous nucleotides. An example of a multi-individual DNA pool is one constructed to obtain population-specific estimates of allele frequency of known polymorphisms ***(42)***. PCR products obtained from such starting material represent more than a single, unique strand of DNA. When these products are sequenced, polymorphic locations between different strands of DNA appear as base ambiguities in the sequence trace (**Fig. 4**). The automation of heterozygote detection motivated the development of POLYPHRED ***(31)***, a computer program ***(43)*** that examines numerical characteristics of sequence traces such as drop in peak-height, ratio of a second peak under the

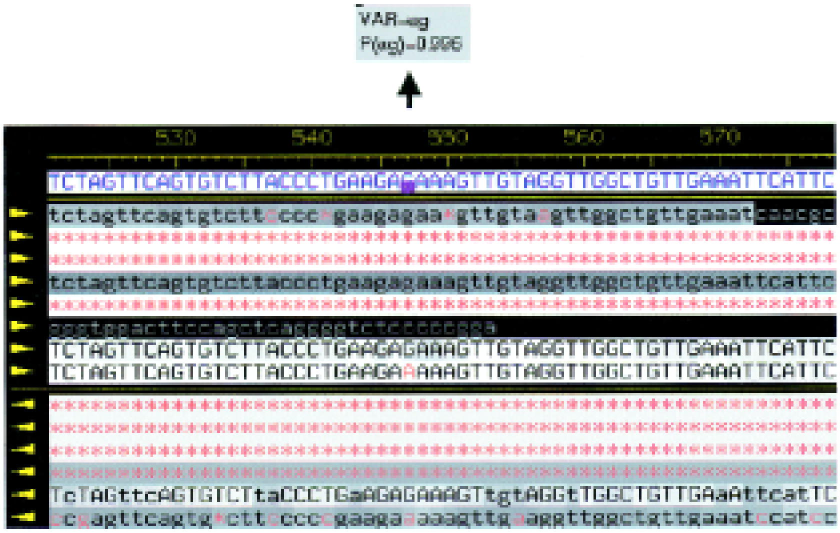

VAR=ag
P(ag)=0.996
530
540
550
560
570
TCTAGTTCAGTGTCTTACCCTGAAGAGAAAGTTGTAGGTTGGCTGTTGAAATTCATTC
tctagttcagtgtcttcccc gaagagaa gttgta gttggctgttgaaat
tctagttcagtgtcttaccctgaagagaaagttgtaggttggctgttgaaattcattc
TCTAGTTCAGTGTCTTACCCTGAAGAGAAAGTTGTAGGTTGGCTGTTGAAATTCATTC
TCTAGTTCAGTGTCTTACCCTGAAGAAAAAGTTGTAGGTTGGCTGTTGAAATTCATTC
TcTAGttcAGTGTCTtaCCCTGaAGAGAAAGTtgtAGGtTGGCTGTTGAaAttcatTC

Fig. 3. Candidate SNP site. The SNP (alleles A/G) is evident within members of one of the two alternatively spliced forms of ESTs aligned to the genomic anchor sequence at this location. The tag above, generated automatically by the detection software POLYBAYES, shows the most likely allele combination at the site, together with the probability of that variation.

primary peak, and overall sequence quality in the neighborhood of the analyzed nucleotide position. POLYHRED integrates seamlessly with the University of Washington PHRED/PHRAP/CONSED genome analysis software package. Although both POLYPHRED, and other specialized, heuristic approaches has been tested for allele frequency estimation in pooled sequencing, reliable computer algorithms of frequency estimation are not yet available.

Another topic of practical importance is the detection of short insertions and deletions (INDELs). Polymorphisms of this type are also commonly referred to as DIPs (deletion-insertion polymorphisms). The main difficulty of detecting DIPs is the fact that current, base-wise measures of sequence accuracy provide no direct

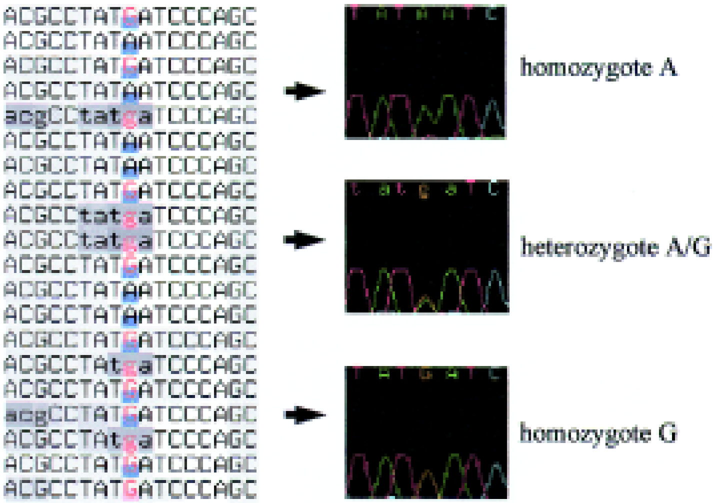


Fig. 4. Heterozygote detection with the POLYPHRED program. Multiple alignment with the site of an SNP marked up with POLYPHRED (left). Sequence traces of a homozygous A/A, a heterozygous A/G, and a homozygous G/G individual (right).

estimates of insertion or deletion type sequencing errors. The base quality value, accompanying a given nucleotide, expresses the likelihood that the nucleotide was called in error, but it is not possible to separate the likelihood of substitution-type sequencing error from the likelihood that a nonexistent nucleotide was artifactually inserted by the base caller. Similarly, there is no direct measure of the likelihood that between two called, neighboring nucleotides there are additional bases in the sequencing template that were erroneously omitted and therefore represent deletion-type errors. In the absence of sequencing error estimates, it is difficult to formulate rigorous models of insertion-deletion type polymorphisms. A heuristic approach employed by POLYBAYES for DIP detection is based on the assumptions that a higher base quality value corresponds to a decreased chance that the called nucleotide is, in fact, an artifactual insertion, and that the likelihood of deleted nucleotides between two high-quality called bases is low. Taking into account the base quality value of the nucleotides neighboring a candidate deletion, as well as the base quality values of the corresponding candidate insertion in another aligned sequence, a heuristic DIP likelihood is calculated. This likelihood was used to detect DIPs in overlapping regions of large-insert clones of the Human Genome Assembly. Validation rate for DIPs that were at least two base pairs long was about 70%; the validation rate for single base-pair insertions-deletions was significantly lower, especially for base-number differences in mono-nucleotide runs.

3.2. Computational Aspects of SNP Discovery

The majority of software packages for automated SNP discovery were developed to run under the UNIX operating system. Part of the reason for this is the availability of powerful and flexible programming tools that UNIX provides for the software developer. In addition, many of the SNP discovery tools available today were written in a way that enables their integration into existing genome analysis packages such as the PHRED/PHRAP/CONSED system, developed at the University of Washington under UNIX. Hardware requirements for SNP mining depend greatly on the scope of the task tackled. Searching for SNPs in specific, short (up to 100–150 kb) regions of the genome, in up to a few hundred sequences, is well within the capabilities of a conventional UNIX workstation (or a computer running the user-friendly LINUX operating system that can be installed on a personal computer with relative ease). Genome-wide SNP mining projects typically require server-class machines, and access to several hundred gigabytes of data storage, especially if intermediate steps of the mining procedure are tracked and results are recorded in a database.

Unfortunately, there is no official standard data exchange format for sequence multiple alignments, or SNP markup information. Many of the SNP discovery tools currently in use expect input and produce output in file formats specific to the program. In these cases, data translation between different tools is achieved via custom scripts. The closest to a *de facto* standard is the PHRED/PHRAP/CONSED ***(24)*** file structure and software architecture developed at the University of Washington that is widely used in sequencing laboratories worldwide. Given that several of the main SNP analysis tools, including POLYPHRED and POLYBAYES, were built to integrate within this structure, it is worthwhile to briefly summarize the University of Washington package standards for representing SNP information.

The main directory of the file architecture contains four subdirectories in which all relevant data is organized. Sequence traces reside in the subdirectory **chromat_dir**. When the base-calling algorithm PHRED interprets a trace, it creates a sequence analysis file in the PHD format, and writes it into the subdirectory **phd_dir**. In addition to header information such as sequence name, read chemistry, and template identifier, the PHD format file contains three important pieces of information for each called base: the called DNA residue, the corresponding base quality value describing the accuracy of the call, and the position of the called nucleotide relative to the sequence trace. The PHD file may also contain permanent additional sequence information or tags attached to sections of the read (such as the region of an annotated repeat, or cloning vector sequence). The pre-requisite of using POLYPHRED is the presence of an additional trace analysis file that contains detailed information about the trace, at the location of the called nucleotide. This file is the POLY format trace analysis file, located in the subdirectory **poly_dir**. Finally, all downstream analysis files are kept in the fourth subdirectory **edit_dir**. Perhaps the most commonly used file in this directory is the ACE format sequence assembly, or multiple alignment file. This file format was designed as an interchange format between the PHRAP sequence assembly program and the CONSED sequence editor. ACE files are versioned and sequence edits performed within CONSED are saved as consecutive versions. The SNP detection program POLYPHRED takes an ace format multiple alignment file, and adds markup information regarding the location of heterozygous trace positions. These tags are visible when the alignment is viewed with CONSED, enabling rapid manual review. POLYBAYES operates in one of two modes. The first mode is the analysis of a pre-existing multiple alignment, supplied in the ACE format. In this case, the anchored multiple alignment step is bypassed, and an ACE format output file is created that contains the results of paralog identification and SNP detection, again, as tags viewable from within CONSED. In the second mode of operation one utilizes the anchored alignment capability of POLYBAYES. In this case, one starts out with FASTA format files representing the DNA sequence and the accompanying base quality values for the genomic anchor sequence, as well as the cluster member sequences (for a description of the FASTA format *see* URL: http://www.ncbi.nlm.nih.gov/BLAST/fasta.html). CROSS_MATCH ***(24)***, a pair-wise, dynamic programming alignment algorithm is run between each member sequence and the anchor. The sequences, together with the pair-wise alignmentsare supplied to POLYBAYES. The program multiply aligns the member sequences, performs the paralog filtering and the SNPdetection step, and produces a new ACE format output file for the viewing of the anchored multiple alignment and SNP analysis results.

3.3. SNP Discovery Protocol

Given the diversity of sequence data that can be used to detect polymorphic sites within an organism, it is impossible to prescribe a single protocol that works in every situation. In general, the mining procedure will contain the following steps: data organization, the creation of a base-wise multiple alignment, filtering of paralogous sequences (or cluster refinement), followed by the detection of SNPs in slices of the multiple alignment. In this final section of this chapter, we will give two different examples that typify the usual steps of SNP mining. The majority of mining applications can be successfully completed by customizing and combining these steps.

3.3.1. SNP Discovery in EST Sequences

In the first scenario, in a screen against a cDNA library one pulls out a clone sequence that contains a gene of interest. The cDNA is an already sequenced clone, the corresponding EST is in the public database, dbEST (37) (URL: http://www.ncbi.nlm.nih.gov/dbEST). The goal is to explore single base-pair variations within the gene. The first step towards this goal is to find all SNPs in those transcribed sequences of the gene that are available in public sequence databases. One proceeds as follows:

1. Find the location of the gene in the human genome from which the EST was expressed. Go to the NCBI (National Center for Biotechnology Information) web site (URL: http://www.ncbi.nlm.nih.gov) and follow the Map Viewer link. Use the search facility on this page to find the genomic location of the EST, pre-computed by the NCBI. Perform the search using the accession number of the EST. Make sure that you set the “Display Settings” to include the “GenBank” view. Click on the genome clone accession that overlaps the EST, and download the sequence in FASTA format. This sequence will act as the genomic anchor sequence for the ESTs to be analyzed.
2. Find all other ESTs in dbEST with significant sequence similarity to the original EST sequence. Perform the similarity search from the NCBI (National Center for Biotechnology Information) website (URL: http://www.ncbi.nlm.nih.gov/BLAST). Choose the “Standard nucleotide-nucleotide BLAST” option. Type the accession number of the EST in the “Search” field. Choose “est_human” as the database to search against. Once the search is done, format the output as “Simple text,” and parse out the accession list of ESTs from the list of hitting sequences (*see* **Note 1**).
3. Retrieve EST sequence traces. In the near future, EST trace retrieval will be possible from the trace repository (URL: http://www.ncbi.nlm.nih.gov/Traces) that is under construction at the NCBI. Currently, EST sequence traces can be downloaded from the Washington University ftp site: (URL: ftp://genome.wustl.edu/pub/gsc1/est) for ESTs produced there. Searching is done via the local EST names. Download all ESTs for which traces can be found at this site (*see* **Note 2**).

4. Process the sequence traces with the PHRED base-calling program. Invoke PHRED with the command line parameters that produce files necessary for downstream processing in the University of Washington PHRED/PHRAP/CONSED architecture (URL: http://www.phrap.org). Make sure that PHD format sequence files are created in the "phd_dir" subdirectory, by specifying the location of this directory with the "-cd" option. Use the utility program PHD2FASTA (provided with CONSED) to produce a FASTA format file of the DNA sequences ("-os" option) of the ESTs file. Also, produce a FASTA format file for the accompanying base quality values ("-oq" option), and one for the list of base positions that specify the location of each called nucleotide relative to the sequence trace ("-ob" option). The DNA sequence of the ESTs will be used in the next step, as the members of the cluster (group) of expressed sequences to analyze for polymorphic sites.
5. Create a multiple alignment of the EST sequences with the anchored alignment algorithm implemented within POLYBAYES (instructions at the POLYBAYES web site, URL: http://genome.wustl.edu/gsc/polybayes). As the anchor sequence, use the genomic clone sequence from **step 1**. Use the CROSS_MATCH dynamic alignment program to compute the initial pair-wise alignments between each of the ESTs and the genomic anchor sequence (CROSS_MATCH is distributed as part of the PHRAP software package ***[24]***). As cluster member sequences, use the ESTs obtained in **steps 2–4**. **Figure 1** shows a section of a sample multiple alignment, viewed with the CONSED ***(36)*** sequence viewer-editor program. Observe that, in this case, the ESTs are divided into two groups of alternative splice forms.
6. Likely paralogous sequences are identified with the in-built paralog-filtering feature of POLYBAYES. This feature is invoked by the "-filterParalogs" command line option (additional relevant arguments explained in the online documentation available at the POLYBAYES web site). **Figure 2** shows a different section of the multiple alignment produced in the previous step. Observe that there are several high-quality mismatches between the genomic anchor sequence and EST marked with the blue tag. This sequence is considered a sequence paralog, and is automatically tagged by the filtering algorithm. The paralogous sequence is removed from consideration in any further analysis.
7. The multiple alignment is scanned for polymorphic sites. At each site, the slice of the alignment composed of nucleotides contributed

by every sequence that was locally aligned, is examined for mismatches. The Bayesian SNP detection algorithm calculates the probability that such mismatches are the result of true polymorphism as opposed to sequencing error. Likely polymorphic sites are recorded as SNP candidates. The SNP detection feature is enabled with the "-screenSnps" option (additional parameters such as setting prior polymorphism rates or the SNP probability threshold, and enabling pre-screening steps, are explained in online the documentation). **Figure 3** shows the site of a SNP candidate in the multiple alignment in the previous example. This SNP is found within members of one alternatively spliced group of EST sequences, and is automatically tagged by the SNP detection algorithm implemented within POLYBAYES (*see* **Note 3**).

A similar procedure is applicable for a wide range of scenarios where sequence fragments (e.g., ESTs, random genomic shotgun reads, BAC-end reads, sequenced restriction fragments, etc.) are organized with the help of genome reference sequence, and compared both against each other, and/or to the reference sequence in search of polymorphic sites.

3.3.2. SNP Discovery in PCR Product Sequences

The second scenario is a genotyping application. The goal is to assay a set of individuals for the presence of polymorphic sites in a small region of interest (such as an exon of a gene). A primer pair is available to amplify the region from genomic DNA. The region is amplified from each individual, and the amplicon sequenced. Whenever an individual is heterozygous for a given allele, the sequence shows an ambiguous (heterozygous) peak. Use POLYPHRED, a software package specifically developed for heterozygote detection, to identify heterozygous positions within sequence traces. The procedure is as follows:

1. Process the sequence traces, each representing the double-stranded, genomic DNA of a single individual, with the PHRED base-calling program. This time, in addition to the trace files and the PHD format sequence files central to the CONSED file structure, also create POLY format trace analysis files. This is done by invoking PHRED

with the "-dd" command line option to specify the **"poly_dir"** subdirectory, within the CONSED structure) where these files are to be written. At the end of this step, a POLY file is present for each of the sequence traces, containing detailed numeric information about the trace characteristics at the position of each called nucleotide.

2. Create a multiple alignment of the sequences representing each of the genotyped individuals. Use the PHRAP fragment assembly program ***(24)*** for this purpose. To enable further analysis of the multiple alignment, invoke PHRAP with the "-new_ace" command line option. This will cause the program to produce an ACE format output file that is suitable for direct analysis by the POLYPHRED program. The ACE format output file can also be directly loaded into the viewer-editor program CONSED for visual review of the multiple alignment.
3. Run POLYPHRED on the multiple alignment to detect polymorphic sites. Using the "-ace" option, specify the "ACE" format PHRAP output file created in the previous step when invoking POLYPHRED. The program analyzes the multiple alignment and tags the sites of candidate SNPs, as identified by likely heterozygous peaks within sequence traces. **Figure 4** shows a section of a multiple alignment containing the site of a SNP, together with examples of sequence traces representing individuals homozygous for each of the two alleles, and a heterozygote.

4. Notes

1. To facilitate the retrieval of the corresponding sequence traces, make a list of local EST read names available in the header information for each EST.
2. The following URL: http://genome.wustl.edu/est/est_search/ftp_guide.html contains detailed instructions.
3. Additional information is provided in the output files produced by the program (for more detail, see the online documentation).

References

1. Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., et al. (2001) A map of human genome

sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–933.

2. Watterson, G. A. (1975) On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7,** 256–276.
3. Halushka, M. K., Fan, J. B., Bentley, K., Hsie, L., Shen, N., Weder, A., et al. (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* **22**, 239–247.
4. Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., et al. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22,** 231–238.
5. Sherry, S. T., Ward, M. H., Kholodov, M., Baker, J., Phan, L., Smigielski, E. M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311.
6. Buetow, K. H., Edmonson, M. N., and Cassidy, A. B. (1999) Reliable identification of large numbers of candidate SNPs from public EST data. *Nat. Genet.* **21**, 323–325.
7. Buetow, K. H., Edmonson, M., MacDonald, R., Clifford, R., Yip, P., Kelley, J., et al. (2001) High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Proc. Natl. Acad. Sci. USA* **98**, 581–584.
8. Irizarry, K., Kustanovich, V., Li, C., Brown, N., Nelson, S., Wong, W., and Lee, C. J. (2000) Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences. *Nat. Genet.* **26**, 233–236.
9. Marth, G. T., Korf, I., Yandell, M. D., Yeh, R. T., Gu, Z., Zakeri, H., et al. (1999) A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* **23,** 452–456.
10. Altshuler, D., Pollara, V. J., Cowles, C. R., Van Etten, W. J., Baldwin, J., Linton, L. and Lander, E. S. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407,** 513-6.
11. Taillon-Miller, P., Gu, Z., Li, Q., Hillier, L., and Kwok, P. Y. (1998) Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms. *Genome Res.* **8**, 748–754.
12. Mullikin, J. C., Hunt, S. E., Cole, C. G., Mortimore, B. J., Rice, C. M., Burton, J., et al. (2000) An SNP map of human chromosome 22. *Nature* **407**, 516–520.

13. Marth, G. T. S., G., Yeh, R., Davenport, R., Agarwala, R., Church, D., Wheelan, S., et al. The structure of single-nucleotide variation in overlapping regions of human genome sequence. In preparation.
14. Fu, Y. X. (1995) Statistical properties of segregating sites. *Theor. Popul. Biol.* **48**, 172–197.
15. Marth, G., Yeh, R., Minton, M., Donaldson, R., Li, Q., Duan, S., et al. (2001) Single-nucleotide polymorphisms in the public domain: how useful are they? *Nat. Genet.* **27**, 371–372.
16. Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., et al. (2001) Linkage disequilibrium in the human genome. *Nature* **411**, 199–204.
17. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
18. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
19. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., et al. (2001) The sequence of the human genome. *Science* **291**, 1304–1351.
20. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680.
21. Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–194.
22. Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8,** 175–185.
23. The Sanger Centre and the Washington University Genome Sequencing Center. T. S. C. a. t. W. U. G. S. (1998) Toward a complete human genome sequence. *Genome Res.* **8**, 1097–1108.
24. Green, P. http://www.phrap.org
25. Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., et al. (2000) A whole-genome assembly of Drosophila. *Science* **287**, 2196–2204.
26. Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., et al. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082.

27. Picoult-Newberg, L., Ideker, T. E., Pohl, M. G., Taylor, S. L., Donaldson, M. A., Nickerson, D. A., and Boyce-Jacino, M. (1999) Mining SNPs from EST databases. *Genome Res.* **9**, 167–174.
28. Garg, K., Green, P., and Nickerson, D. A. (1999) Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. *Genome Res.* **9**, 1087–1092.
29. Hillier, L. D., Lennon, G., Becker, M., Bonaldo, M. F., Chiapelli, B., Chissoe, S., et al. (1996) Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* **6**, 807–828.
30. Adams, M. D., Soares, M. B., Kerlavage, A. R., Fields, C., and Venter, J. C. (1993) Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nat. Genet.* **4**, 373–380.
31. Nickerson, D. A., Tobe, V. O., and Taylor, S. L. (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* **25**, 2745–2751.
32. Dawson, E., Chen, Y., Hunt, S., Smink, L. J., Hunt, A., Rice, K., et al. (2001) A SNP resource for human chromosome 22: extracting dense clusters of SNPs from the genomic sequence. *Genome Res.* **11**, 170–178.
33. Kwok, P.-Y., Carlson, C., Yager, T. D., Ankener, W., and Nickerson, D. A. (1994) Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products. *Genomics* **23**, 138–144.
34. Nickerson, D. A., Taylor, S. L., Weiss, K. M., Clark, A. G., Hutchinson, R. G., Stengard, J., et al. (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nat. Genet.* **19**, 233–240.
35. Nickerson, D. A., Taylor, S. L., Fullerton, S. M., Weiss, K. M., Clark, A. G., Stengard, J. H., et al. (2000) Sequence diversity and large-scale typing of SNPs in the human apolipoprotein E gene. *Genome Res.* **10**, 1532–1545.
36. Gordon, D., Abajian, C., and Green, P. (1998) Consed: a graphical tool for sequence finishing. *Genome Res.* **8**, 195–202.
37. Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M. (1993) dbEST: database for “expressed sequence tags”. *Nat. Genet.* **4**, 332–333.
38. Wheeler, D. L., Church, D. M., Lash, A. E., Leipe, D. D., Madden, T. L., Pontius, J. U., et al. (2001) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **29**, 11–16.

39. Smit, A. F. A. G., P., http://ftp.genome.washington.edu/RM/RepeatMasker.html
40. Ning, Z., Cox, A. J., and Mullikin, J. C. (2001) SSAHA: A fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729.
41. Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., and Walters, L. (1998) New goals for the U.S. Human Genome Project: 1998–2003. *Science* **282**, 682–689.
42. Kwok, P.-Y. (2000) Approaches to allele frequency determination. *Pharmacogenomics* **1**, 231–235.
43. Nickerson, D. A., http://droog.mbt.washington.edu/PolyPhred.html